Optimal Training in Space-Time Systems

Babak Hassibi and Bertrand Hochwald
Mathematics of Communications Research
Bell Labs, Lucent Technologies, Murray Hill, NJ 07974

Abstract

Multiple-antenna wireless communication links promise very high data rates with low error probabilities, especially when the wireless channel response is known at the receiver. In practice, knowledge of the channel is often obtained by sending known training symbols to the receiver. We show how training affects the capacity of a fading channel—too little training and the channel is improperly learned, too much training and there is no time left for data transmission before the channel changes. We use an information-theoretic approach to compute the optimal amount of training as a function of the received signal-to-noise ratio, fading coherence time, and number of transmitter antennas. When the training and data powers are allowed to vary, we show that the optimal number of training symbols is equal to the number of transmit antennas—this number is also the smallest training interval length that guarantees meaningful estimates of the channel matrix. When the training and data powers are instead required to be equal, the optimal number of symbols may be larger than the number of antennas. We further conclude that at high SNR training-based schemes can capture most of the channel capacity, whereas at low SNR they can be highly suboptimal.

1 Introduction

Multiple-antenna wireless communication links promise very high data rates with low error probabilities, especially when the wireless channel response is known at the receiver [1, 2]. To learn the channel, the receiver often requires the transmitter to send known training signals during some portion of the transmission interval. In this paper we address the following problem: given a multi-antenna wireless link with M transmit antennas, N receive antennas, coherence interval of length T (in symbols), and SNR ρ, *how much of the coherence interval should be spent training?*

Our solution is based on a lower bound on the information-theoretic capacity achievable with training-based schemes. The lower bound allows us to compute the optimal amount of training as a function of ρ, T, M, and N. We are also able to identify some occasions where training imposes a substantial information-theoretic penalty, especially at low SNR or when the coherence interval T is only slightly larger than the number of transmit antennas M. Conversely, if the SNR is high and T is much larger than M, then training-based schemes can come very close to achieving capacity.

2 Channel Model and Problem Statement

We assume that the channel obeys the simple discrete-time *block-fading* law, where the channel is constant for some discrete time interval T, after which it changes to an independent value which it holds for another interval T, and so on. We further assume that channel estimation (via training) and data transmission is to be done within the interval T, after which new training allows us to estimate the channel for the next T symbols, and so on.

Within one block of T symbols, the multiple-antenna model is

$$X = \sqrt{\frac{\rho}{M}}SH + V, \tag{1}$$

where X is a $T \times N$ received complex signal matrix, the dimension N representing the number of receive antennas. The transmitted signal is S, a $T \times M$ complex matrix where M is the number of transmit antennas. The $M \times N$ matrix H represents the channel connecting the M transmit to the N receive antennas, and V is a $T \times N$ matrix of additive noise. The matrices H and V both comprise independent zero-mean unit-variance complex-Gaussian entries. We also assume that the entries of the transmitted signal S have unit mean-square. Thus, ρ is the expected received SNR at each receive antenna.

0-7803-6514-3/00/$10.00©2000 IEEE

2.1 Training-based schemes

Training-based schemes are composed of the following two phases.

1. **Training Phase:** Here we may write

$$X_\tau = \sqrt{\frac{\rho_\tau}{M}} S_\tau H + V_\tau,\ S_\tau \in \mathcal{C}^{T_\tau \times M},\ \mathrm{tr} S_\tau S_\tau^* = M T_\tau \tag{2}$$

where S_τ is the matrix of training symbols sent over T_τ time samples and known to the receiver, and ρ_τ is the SNR during the training phase. The observed signal matrix $X_\tau \in \mathcal{C}^{T_\tau \times N}$ and S_τ are used to construct an estimate of the channel

$$\hat{H} = f(X_\tau, S_\tau). \tag{3}$$

To obtain a meaningful estimate of H, we need at least as many measurements as unknowns, which implies that $N \cdot T_\tau \geqslant N \cdot M$ or $T_\tau \geqslant M$.

2. **Data Transmission Phase:** Here we may write

$$X_d = \sqrt{\frac{\rho_d}{M}} S_d H + V_d,\ S_d \in \mathcal{C}^{T_d \times M},\ \mathrm{E\,tr} S_d S_d^* = M T_d \tag{4}$$

where S_d is the matrix of data symbols sent over T_d time samples, ρ_d is the SNR during the data transmission phase, and $X_d \in \mathcal{C}^{T_d \times N}$ is the received matrix. The estimate of the channel $\hat{H}$ is used to recover S_d. This is written formally as

$$X_d = \sqrt{\frac{\rho_d}{M}} S_d \hat{H} + \underbrace{\sqrt{\frac{\rho_d}{M}} S_d \tilde{H} + V_d}_{V_d'}, \tag{5}$$

where $\tilde{H} = H - \hat{H}$ is the channel estimation error.

This two-phase training and data process is equivalent to partitioning the matrices in (1) as

$$S = \begin{pmatrix} \sqrt{\frac{\rho_\tau}{\rho}} S_\tau \\ \sqrt{\frac{\rho_d}{\rho}} S_d \end{pmatrix},\ X = \begin{pmatrix} X_\tau \\ X_d \end{pmatrix},\ V = \begin{pmatrix} V_\tau \\ V_d \end{pmatrix}.$$

Conservation of time and energy yield

$$T = T_\tau + T_d, \qquad \rho T = \rho_\tau T_\tau + \rho_d T_d. \tag{6}$$

3 Capacity and Capacity Bounds

Strictly speaking, as long the estimate of the channel matrix $\hat{H} = f(X_\tau, S_\tau)$ does not "throw away" information, the choice of the channel estimate in (5) does not affect the capacity because the capacity depends only on the conditional distribution of H given S_τ and X_τ. But most practical data transmission schemes that employ training do throw away information because they use the estimate $\hat{H}$ as if it were correct. We assume that such a scheme is employed.

In particular, we find a lower bound on the capacity by choosing a particular estimate of the channel. We assume that $\hat{H}$ is the conditional mean of H, given S_τ and X_τ. By well-known properties of the conditional mean, $\hat{H}$ and $\tilde{H}$ are uncorrelated.

The channel model during the data transmission phase is given by (4), where V_d' combines the additive noise and residual channel estimation error. The estimate $\hat{H} = f(X_\tau, S_\tau)$ is known and assumed by the training-based scheme to be correct; hence, the channel capacity of a training-based scheme is the same as the capacity of a *known channel* system, subject to additive noise with the power constraint

$$\sigma_{v'}^2 = \frac{1}{NT_d} \mathrm{trE}\, V_d' V_d'^* = 1 + \rho_d \sigma_{\tilde{H}}^2,$$

where $\sigma_{\tilde{H}}^2 \triangleq \frac{1}{NM} \mathrm{E\,tr} \tilde{H}\tilde{H}^*$. There are two important differences between (4) and (1). In (4) the channel is known to the receiver whereas in (1) it is not. In (1) the additive noise is Gaussian and independent of the data whereas in (4) it is possibly neither. Since we have used the conditional mean as the estimate $\hat{H}$, all that we know about V_d' is that it is uncorrelated with the signal S_d. Thus, finding the capacity of a training-based scheme requires us to examine the worst effect the *uncorrelated* additive noise can have during data transmission. This is the content of the next theorem, which is proven in [3].

Theorem 1 (Worst-Case Uncorrelated Additive Noise) *Consider the matrix-valued additive noise channel*

$$X = \sqrt{\frac{\rho}{M}} SH + V,$$

where $H \in \mathcal{C}^{M \times N}$ is the known channel, and where the signal $S \in \mathcal{C}^{1 \times M}$ and the additive noise $V \in \mathcal{C}^{1 \times N}$ satisfy the power constraints $Exp \frac{1}{M} SS^* = 1$ and $\mathrm{E}\frac{1}{N} VV^* = 1$, and are uncorrelated: $\mathrm{E}\, S^* V = 0_{M \times N}$. Then the worst-case noise has an iid zero-mean Gaussian distribution, $V \sim \mathcal{CN}(0, I_N)$.

The noise term V_d' in (4), when $\hat{H}$ is the MMSE estimate, is uncorrelated with S_d but is not necessarily Gaussian. Theorem 1 says that a lower bound on the training-based capacity is obtained by replacing V_d' by independent zero-mean spatially and temporally white additive Gaussian noise of the same variance $1 + \rho_d \sigma_{\tilde{H}}^2$. Using

the formula for the capacity in an additive Gaussian noise known channel [1, 2], we may therefore write

$$C_\tau \geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\log\det\left(I_M+\frac{\rho_d}{1+\rho_d\sigma^2_{\hat{H}}}\frac{\hat{H}\hat{H}^*}{M}\right),$$

where C_τ denotes the training-based capacity, and the coefficient $T-T_\tau$ reflects the fact that the data transmission phase has a duration of $T_d = T-T_\tau$ time symbols. Since $\hat{H}$ is zero-mean its variance can be defined as $\sigma^2_{\hat{H}} = \frac{1}{NM}\mathrm{E}\,\mathrm{tr}\hat{H}\hat{H}^*$. By the orthogonality principle for MMSE estimates, $\sigma^2_{\hat{H}} = 1-\sigma^2_{\tilde{H}}$, and we can define the *normalized channel estimate* as $\bar{H} \triangleq \frac{1}{\sigma_{\hat{H}}}\hat{H}$. We may thus write the capacity bound as

$$C_\tau \geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\log\det\left(I_M+\frac{\rho_d\sigma^2_{\hat{H}}}{1+\rho_d\sigma^2_{\tilde{H}}}\frac{\bar{H}\bar{H}^*}{M}\right). \tag{7}$$

The ratio $\rho_{\text{eff}} = \frac{\rho_d\sigma^2_{\hat{H}}}{1+\rho_d\sigma^2_{\tilde{H}}}$ can be considered as an *effective* SNR. The remainder of this paper is concerned with maximizing this lower bound bu choosing, (I) the training data S_τ, (II) the training power ρ_τ, and (III) the training interval length T_τ.

Finally, we note that (7) is a bound since $V'_d = \sqrt{\rho_d/M}S_d\tilde{H}+V_d$ is not, in general, Gaussian. However, since V_d is Gaussian, V'_d becomes Gaussian as $\rho_d \to 0$, and so the bound (7) becomes tight in this regime. In Section 3.3.1 we use this tightness to conclude that training is suboptimal at low SNR. In Section 5 we show that this bound is also tight at high SNR. We therefore expect this bound to be reasonably tight for a wide range of SNR's.

3.1 Optimizing over S_τ

From (7) it is clear that S_τ affects the capacity bound only through the effective SNR:

$$\rho_{\text{eff}} = \frac{\rho_d\sigma^2_{\hat{H}}}{1+\rho_d\sigma^2_{\tilde{H}}} = \frac{\rho_d(1-\sigma^2_{\tilde{H}})}{1+\rho_d\sigma^2_{\tilde{H}}} = \frac{1+\rho_d}{1+\rho_d\sigma^2_{\tilde{H}}} - 1.$$

It therefore follows that we need to choose S_τ to minimize the mean-square-error $\sigma^2_{\tilde{H}} = \frac{1}{M}\mathrm{tr}\left(I_M+\frac{\rho_\tau}{M}S^*_\tau S_\tau\right)^{-1}$. This is solved by setting $S^*_\tau S_\tau = T_\tau I_M$, as the optimal training signal; i.e., *the training signal must be a multiple of a matrix with orthonormal columns.*

With this choice of training signal, it follows that

$$C_\tau \geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\log\det\left(I_M+\rho_{\text{eff}}\frac{\bar{H}\bar{H}^*}{M}\right), \tag{8}$$

where

$$\rho_{\text{eff}} = \frac{\rho_d\rho_\tau T_\tau}{M(1+\rho_d)+\rho_\tau T_\tau}, \tag{9}$$

and where $\bar{H}$ has independent $\mathcal{CN}(0,1)$ entries.

3.2 Optimizing over the power allocation

The power allocation $\{\rho_d, \rho_\tau\}$ enters the capacity formula via ρ_{eff} only. Thus, we need to choose $\{\rho_d, \rho_\tau\}$ to maximize ρ_{eff}.

Theorem 2 (Optimal Power Distribution). *The optimal power allocation $\alpha = \frac{\rho_d T_d}{\rho T}$ in a training-based scheme is given by*

$$\alpha = \begin{cases} \gamma-\sqrt{\gamma(\gamma-1)} & \text{for } T_d > M \\ \frac{1}{2} & \text{for } T_d = M \\ \gamma+\sqrt{\gamma(\gamma-1)} & \text{for } T_d < M \end{cases} \tag{10}$$

where $\gamma = \frac{M+\rho T}{\rho T(1-\frac{M}{T_d})}$. The corresponding capacity lower bound is

$$C_\tau \geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\log\det\left(I_M+\rho_{\text{eff}}\frac{\bar{H}\bar{H}^*}{M}\right), \tag{11}$$

where

$$\rho_{\text{eff}} = \begin{cases} \frac{\rho T}{T_d-M}\left(\sqrt{\gamma}-\sqrt{\gamma-1}\right)^2 & \text{for } T_d > M \\ \frac{(\rho T)^2}{4M(M+\rho T)} & \text{for } T_d = M \\ \frac{\rho T}{M-T_d}\left(\sqrt{-\gamma}-\sqrt{-\gamma+1}\right)^2 & \text{for } T_d < M \end{cases} \tag{12}$$

These formulas are especially revealing at high and low SNR.

Corollary 1 (High and Low SNR). *1. At high SNR*

$$\alpha = \frac{\sqrt{T_d}}{\sqrt{T_d}+\sqrt{M}}, \qquad \rho_{\text{eff}} = \frac{T}{(\sqrt{T_d}+\sqrt{M})^2}\rho.$$

2. At low SNR: $\alpha = \frac{1}{2}$ and $\rho_{\text{eff}} = \frac{T^2}{4MT_d}\rho^2$.

3.3 Optimizing over T_τ

All that remains is to determine the length of the training interval T_τ. We show that setting $T_\tau = M$ is optimal for any ρ and T (provided that we optimize ρ_τ and ρ_d). There is a simple intuitive explanation for this result. Increasing T_τ beyond M linearly decreases the capacity through the $\frac{T-T_\tau}{T}$ term in (11), but only logarithmically increases the capacity through the higher effective SNR ρ_{eff}. We therefore have a natural tendency to make T_τ as

small as possible. Although making T_τ small loses accuracy in estimating H, we can compensate for this loss by increasing ρ_τ (even though this decreases ρ_d). We have the following result, which is the last step in our list of optimizations.

Theorem 3 (Optimal Training Interval). *The optimal length of the training interval is $T_\tau = M$ for all ρ and T, and the capacity lower bound is*

$$C_\tau \geq \mathrm{E}\,\frac{T-M}{T}\log\det\left(I_M + \rho_{\mathrm{eff}}\frac{\bar{H}\bar{H}^*}{M}\right), \quad (13)$$

where

$$\rho_{\mathrm{eff}} = \begin{cases} \frac{\rho T}{T-2M}(\sqrt{\gamma}-\sqrt{\gamma-1})^2 & \text{for } T > 2M \\ \frac{\rho^2}{1+2\rho} & \text{for } T = 2M \\ \frac{\rho T}{2M-T}(\sqrt{-\gamma}-\sqrt{-\gamma+1})^2 & \text{for } T < 2M \end{cases} \quad (14)$$

The optimal allocation of power is as given in (10) with $T_d = T - T_\tau = T - M$ and can be approximated at high SNR by

$$\alpha_{opt} = \frac{\sqrt{T-M}}{\sqrt{T-M}+\sqrt{M}},\ \rho_{\mathrm{eff}} = \frac{1}{(\sqrt{1-\frac{M}{T}}+\sqrt{\frac{M}{T}})^2}\rho \quad (15)$$

and the power allocation becomes

$$\rho_d = \frac{\rho}{1-\frac{M}{T}+\sqrt{(1-\frac{M}{T})\frac{M}{T}}},\ \rho_\tau = \frac{\rho}{\frac{M}{T}+\sqrt{(1-\frac{M}{T})\frac{M}{T}}}. \quad (16)$$

3.3.1 Low SNR

At low SNR the capacity is actually not sensitive to the length of the training interval. From (11) and (12), for small ρ, we may write

$$(\sqrt{\gamma}-\sqrt{\gamma-1})^2 \approx \frac{\rho T(T_d-M)}{4MT_d}$$

to obtain

$$\begin{aligned} C_\tau &\geq \frac{T_d}{T}\mathrm{E}\,\mathrm{tr}\,\log\left(I_M + \frac{T^2}{4MT_d}\rho^2\frac{\bar{H}\bar{H}^*}{M}\right) \\ &\approx \frac{T_d}{T}(\log e)\mathrm{E}\,\mathrm{tr}\left(\frac{T^2}{4MT_d}\rho^2\frac{\bar{H}\bar{H}^*}{M}\right) = \frac{NT\log e}{4M}\rho^2, \end{aligned}$$

which is independent of T_τ.

Note also that at low SNR the capacity with training decays as ρ^2. However, the true channel capacity (which does not necessarily require training to achieve) decays as ρ for small ρ [4]. We therefore must conclude that training is highly suboptimal when ρ is small.

3.4 Equal training and data power

A communication system often does not have the luxury of varying the power during the training and data phases. If we assume that the training and data symbols are transmitted at the same power $\rho_\tau = \rho_d = \rho$ then (8) and (9) become

$$C_\tau \geq \mathrm{E}\,\frac{T-T_\tau}{T}\log\det\left(I_M + \frac{\rho^2 T_\tau/M}{1+(1+T_\tau/M)\rho}\frac{\bar{H}\bar{H}^*}{M}\right) \quad (17)$$

Here the optimal T_τ depends on ρ, T, M, and N, and can be obtained by evaluating the lower bound in (17) (either analytically, see, e.g., [1], or via Monte Carlo simulation) for various values of T_τ.

4 Simulation Results

Figure 1 displays the capacity obtained as a function of the blocklength T for $M = N = 10$ when ρ_τ and ρ_d are optimized versus $\rho_\tau = \rho_d = \rho$.

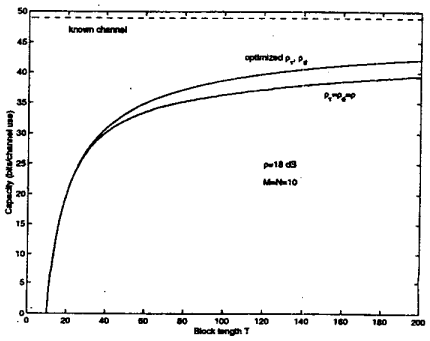

Figure 1. The training-based lower bound on capacity as a function of T when SNR $\rho = 18$ dB and $M = N = 10$, for optimized ρ_τ and ρ_d, and for $\rho_\tau = \rho$.

Figure 2 displays the T_τ that maximizes (17) for different values of ρ with $M = N = 10$. We see the trend that as the SNR decreases, the amount of training increases.

For a given SNR ρ, coherence interval T, and number of receive antennas N, we can calculate the capacity lower bound as a function of M. For $M \approx 1$, the training-based capacity is small because there are few antennas, and for $M \approx T$ the capacity is again small because we spend the entire coherence interval training. We can seek the value of M that maximizes this capacity. Figure 3 shows the capacity as a function of M for

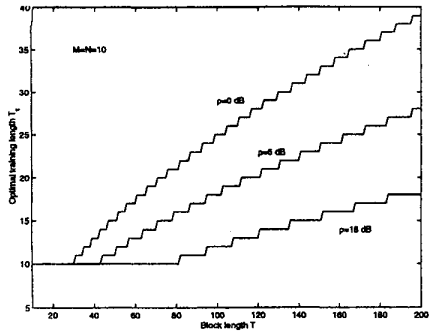

Figure 2. The optimal amount of training T_τ as a function of block length T for for $M = N = 10$ and $\rho_\tau = \rho_d = \rho$.

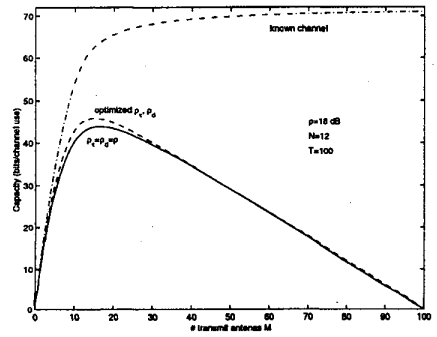

Figure 3. Capacity as a function of number of transmit antennas M with $\rho = 18$ dB and $N = 12$ receive antennas.

$\rho = 18$ dB, $N = 12$, and $T = 100$. We see that choosing to train with the wrong number of antennas can severely hurt the data rate.

5 Discussion and Conclusion

The lower bounds on the capacity of multiple-antenna training-based schemes show that optimizing over the power allocation ρ_τ and ρ_d makes the optimum length of the training interval T_τ equal to M for all ρ and T. If we require the power allocation for training and transmission to be the same, then the length of the training interval can be longer than M, although simulations at high SNR suggest that it is not much longer. As the SNR decreases, however, the training interval increases until at low SNR it converges to half the coherence interval.

The lower bounds on the capacity suggest that training-based schemes are highly suboptimal when T is "close" to M. Figure 3 suggests that it is beneficial to use a training-based scheme with a smaller number of antennas $K < M$. We may ask what is the optimal value of K? To answer this, we suppose that M antennas are available but we elect to use only $K \leqslant M$ of them in a training-based scheme. Thus: as

$$C(\rho, T, M, N) \geqslant \max_{K \leqslant M} C_\tau(\rho, T, K, N). \qquad (18)$$

Using Theorem 3, it is not difficult to show that at high SNR the optimal value is $K = \min(M, N, T/2)$ and that

$$C(\rho, T, M, N) \geqslant \left(1 - \frac{K}{T}\right) K \log \rho. \qquad (19)$$

We argued in Section 3 that the whole process of training is highly suboptimal at low SNR. We now ask whether the same is true at high SNR, and whether our bounds are tight? The answer to this question can be found in the recent work [5] of Zheng and Tse where it is shown that at high SNR the leading term of the actual channel capacity (*without* imposing any constraints such as training) is $\left(1 - \frac{K}{T}\right) K \log \rho$. Thus, in the leading SNR term (as $\rho \to \infty$), training-based schemes are optimal, provided we use $K = \min(M, N, T/2)$ transmit antennas.

References

[1] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecom.*, vol. 10, pp. 585–595, Nov. 1999.

[2] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs. Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[3] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[4] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, "The capacity of discrete-time Rayleigh fading channels," in *IEEE Int. Symp. Info. Theory*, p. 473, June 1997. Also submitted to *IEEE Trans. Info. Theory*.

[5] L. Zheng and D. Tse, "Packing spheres in the Grassman manifold: a geometric approach to the noncoherent multi-antenna channel," *submitted to IEEE Trans. Info. Theory*, 2000.